Large-Scale Performance Analysis Using the BIPS Application Benchmark Suite

Leonid Oliker
Computational Research Division
Lawrence Berkeley National Laboratory

Overview

- Stagnating application performance is well-know problem in scientific computing
- By end of decade numerous mission critical applications expected to have 100X computational demands of current levels
- Many HEC platforms are poorly balanced for demands of leading applications
  - Memory-CPU gap, deep memory hierarchies, poor network-processor integration, low-degree network topology
- Traditional superscalar trends slowing down
  - Mined most benefits of ILP and pipelining, Clock frequency limited by power concerns
- In order to continuously increase computing power <u>and</u> reap its benefits: major strides necessary in architecture development, software infrastructure, and application development

Application Evaluation

- Microbenchmarks, algorithmic kernels, performance modeling and prediction, are important components of understanding and improving architectural efficiency
- However full-scale application performance is the final arbiter of system utility and necessary as baseline to support all complementary approaches
- Our evaluation work emphasizes full applications, with real input data, at the appropriate scale
- Requires coordination of computer scientists and application experts from highly diverse backgrounds
- Our initial efforts have focused on comparing performance between high-end vector and scalar platforms
- Effective code vectorization is an integral part of the process
    - First US team to conduct Earth Simulator performance study

Benefits of Evaluation

BIPS

- Full scale application evaluation lead to more efficient use of the community resources
  - For both current installation and future designs
- Head-to-head comparisons on full applications:
  - Help identify the suitability of a particular architecture for a given application class
  - Give application scientists information about how well various numerical methods perform across systems
  - Reveal performance-limiting system bottlenecks that can aid designers of the next generation systems.
    - Science Driven Architecture
- In-depth studies reveal limitation of compilers, operating systems, and hardware, since all of these components must work together at scale to achieve high performance.

Application Overview

Examining set of applications with potential to run at ultra-scale and abundant data parallelism

NAME	Discipline	Problem/Method	Structure
MADCAP	Cosmology	CMB analysis	Dense Matrix
CACTUS	Astrophysics	Theory of GR	Grid
LBMHD	Plasma Physics	MHD	Lattice
GTC	Magnetic Fusion	Vlasov-Poisson	Particle
PARATEC	Material Science	DFT	Fourier/Grid
FVCAM	Climate Modeling	AGCM	Grid
SuperNova	*Combustion*	*Rayleigh-Taylor*	*AMR Grid*
SuperLU	*Linear Algebra*	*Sparse Direct LU*	*Sparse Matrix*
PMEMD	*Life Sciences*	*Particle Mesh Ewald*	*Particle*

Architectural Comparison

Node Type	Where	Network	CPU/ Node	Clock MHz	Peak GFlop	Stream BW GB/s/P	Peak byte/flop	MPI BW GB/s/P	MPI Latency μsec	Network Topology
Power3	NERSC	Colony	16	375	1.5	0.4	0.26	0.13	16.3	Fat-tree
Itanium2	LLNL	Quadrics	4	1400	5.6	1.1	0.19	0.25	3.0	Fat-tree
Opteron	NERSC	InfiniBand	2	2200	4.4	2.3	0.51	0.59	6.0	Fat-tree
X1	ORNL	Custom	4	800	12.8	14.9	1.16	6.3	7.1	4D-Hypercube
X1E	ORNL	Custom	4	1130	18.0	9.7	0.54	2.9	5.0	4D-Hypercube
ES	ESC	IN	8	1000	8.0	26.3	3.29	1.5	5.6	Crossbar
SX-8	HLRS	INX	8	2000	16.0	41.0	2.56	2.0	5.0	Crossbar

- Custom vector architectures have
  - High memory bandwidth relative to peak
  - Superior interconnect: latency, point to point, and bisection bandwidth
- Overall ES appears as the most balanced architecture
- Jacquard (Opteron/IB) best balance for superscalar arch, Thunder (Itanium2/Quadrics) lowest latency
- A key 'balance point' for vector systems is the scalar:vector ratio

adxsd

IPM Overview

Integrated Performance Monitoring

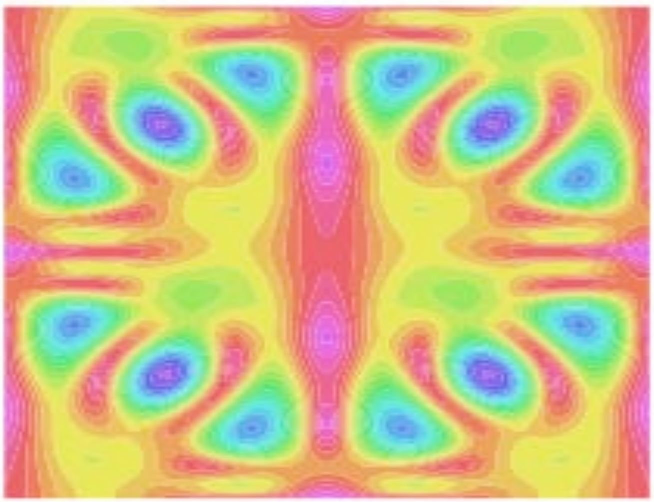
- portable, lightweight, scalable profiling
- fast hash method
- profiles MPI topology
- profiles code regions
- open source

```
MPI_Pcontrol(1,"W");
 …code…
MPI_Pcontrol(-1,"W");
```

```
############################################
# IPMv0.7 :: csnode041 256 tasks  ES/ESOS
# madbench.x (completed) 10/27/04/14:45:56
#
#        <mpi>      <user>      <wall> (sec)
#       171.67      352.16      393.80
# …
#############################################
# W
#        <mpi>      <user>      <wall> (sec)
#        36.40      198.00      198.36
#
# call              [time]       %mpi   %wall
# MPI_Reduce        2.395e+01    65.8     6.1
# MPI_Recv          9.625e+00    26.4     2.4
# MPI_Send          2.708e+00     7.4     0.7
# MPI_Testall       7.310e-02     0.2     0.0
# MPI_Isend         2.597e-02     0.1     0.0
#############################################
…
```

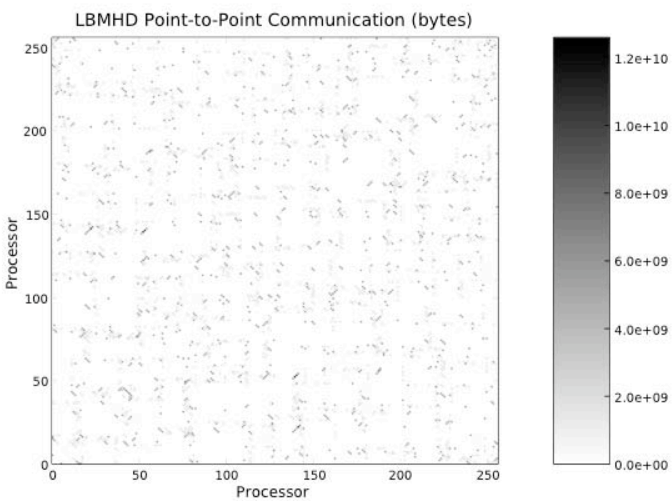
Developed by David Skinner, LBNL

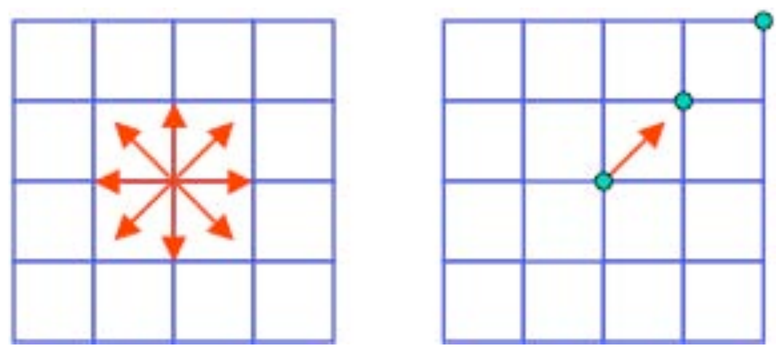
Plasma Physics: LBMHD



Evolution of vorticity into turbulent structures

- LBMHD uses a Lattice Boltzmann method to model magneto-hydrodynamics (MHD)
- Performs 2D/3D simulation of high temperature plasma
- Evolves from initial conditions and decaying to form current sheets
- Spatial grid is coupled to octagonal streaming lattice
- Block distributed over processor grid





Developed by George Vahala's group College of William & Mary, ported Jonathan Carter

BIPS

LBMHD-3D: Performance

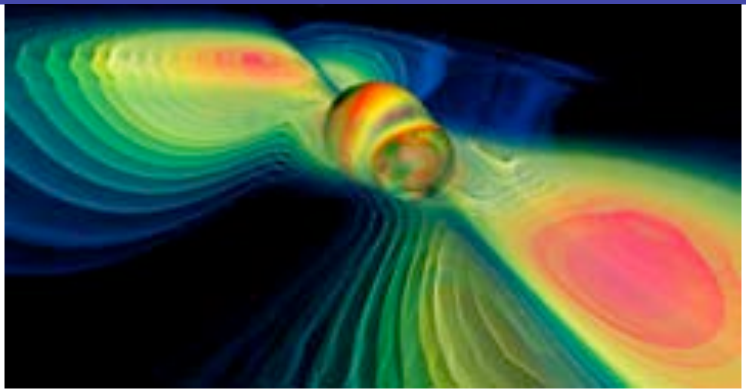
BERKELEY LAB

Grid Size	P	Power3 Seaborg		Itanium2 Thunder		Opteron Jacquard		X1 Phoenix		X1E Phoenix		SX6 ES		SX8 HLRS	
		GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk
256^3	16	0.14	9%	0.26	5%	0.70	16%	5.2	41%	6.6	37%	5.5	69%	7.9	49%
512^3	64	0.15	9%	0.35	6%	0.68	15%	5.2	41%	5.8	32%	5.3	66%	8.1	51%
1024^3	256	0.14	9%	0.32	6%	0.60	14%	5.2	41%	6.0	33%	5.5	68%	9.6	60%
2048^3	512	0.14	9%	0.35	6%	0.59	13%			5.8	32%	5.2	65%		

- Not unusual to see vector achieve > 40% peak while superscalar architectures achieve < 10%
- There exists plenty of computation, however large working set causes register spilling scalars
- Opteron shows impressive superscalar performance
  - Itanium2 has same memory bandwidth as Opteron but cannot store FP in L1
- Large vector register sets hide latency
- ES sustains 68% of peak up to 4800 processors: 26TFlops - the highest performance ever attained for this code by far!
- SX8 shows highest raw performance, but lags behind ES in terms of efficiency
  - SX8: Commodity DDR2-SDRAM vs. ES: high performance custom FPLRAM
- X1E achieved same performance as X1 using original code version
  - By turning off caching resulted in about 10% improvement over X1

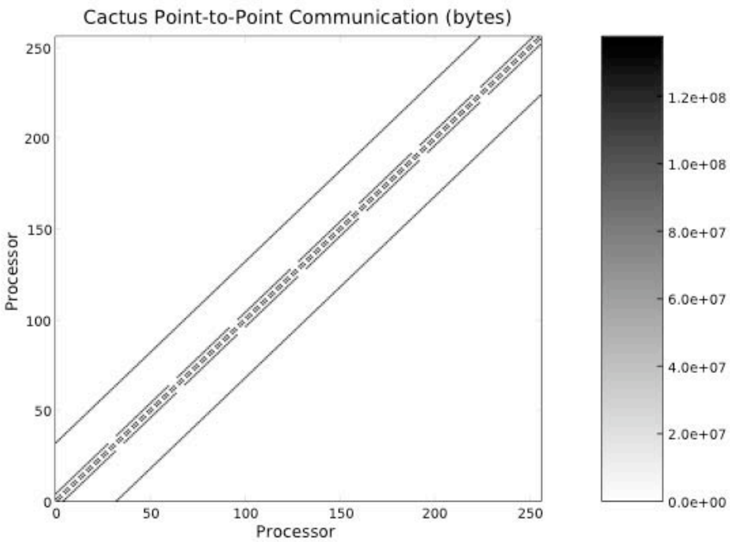
Office of Science
U.S. DEPARTMENT OF ENERGY

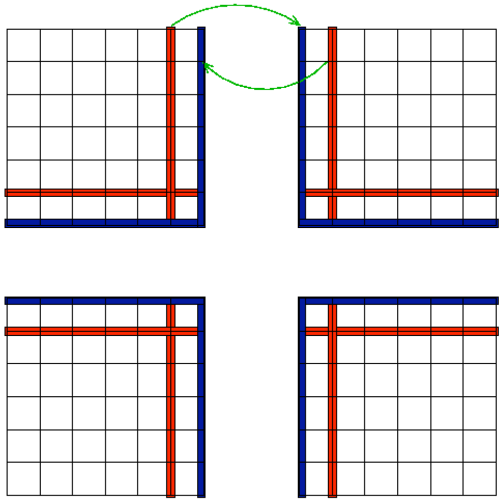
Astrophysics: CACTUS



Visualization of grazing collision of two black holes

- Numerical solution of Einstein's equations from theory of general relativity
- Among most complex in physics: set of coupled nonlinear hyperbolic & elliptic systems with thousands of terms
- CACTUS evolves these equations to simulate high gravitational fluxes, such as collision of two black holes
- Evolves PDE's on regular grid using finite differences





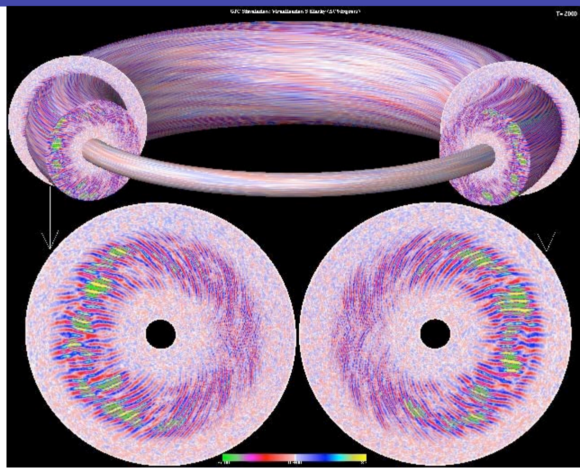
Developed at Max Planck Institute, vectorized by John Shalf LBNL

CACTUS: Performance

Problem Size	P	NERSC (Power 3)		Thunder (Itan2)		Phoenix (X1)		ES (SX6*)		SX8	
		GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk
250x80x80 per processor	16	0.10	6%	0.58	10%	0.81	6%	2.8	35%	4.3	27%
	64	0.08	6%	0.56	10%	0.72	6%	2.7	34%		
	256	0.07	5%	0.55	10%	0.68	5%	2.7	34%		

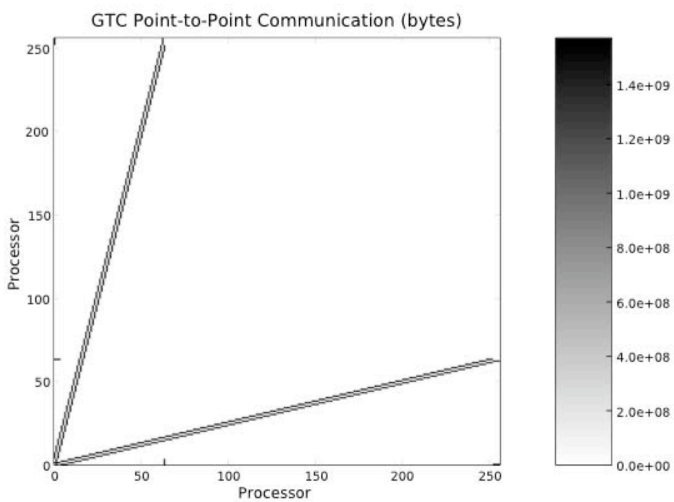
- SX8 attains highest per-processor performance ever attained for Cactus
- ES achieves highest overall performance and efficiency to date: 39X faster than Power3!
  - Vector performance related to x-dim (vector length)
  - Excellent scaling on ES using fixed data size per proc (weak scaling)
  - Opens possibility of computations at unprecedented scale
- X1 surprisingly poor (4X slower than ES) - low ratio scalar:vector
  - Unvectorized boundary, required 15% of runtime on ES and 30+% on X1
  - < 5% for the scalar version: **unvectorized code can quickly dominate cost**
- Poor superscalar performance despite high computational intensity
  - Register spilling due to large number of loop variables
  - Prefetch engines inhibited due to multi-layer ghost zones calculations

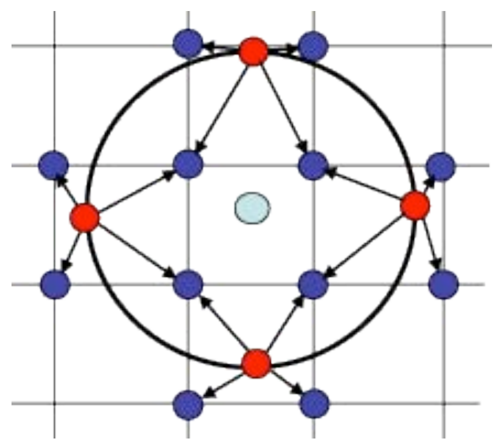
Magnetic Fusion: GTC



Electrostatic potential in magnetic fusion device

- Gyrokinetic Toroidal Code: transport of thermal energy (plasma microturbulence)
- Goal magnetic fusion is burning plasma power plant producing cleaner energy
- GTC solves 3D gyroaveraged gyrokinetic system w/ particle-in-cell approach (PIC)
- PIC scales N instead of N^2 – particles interact w/ electromagnetic field on grid
- Allows solving equation of particle motion with ODEs (instead of nonlinear PDEs)





Developed at Princeton Plasma Physics Laboratory, vectorized by Stephane Ethier

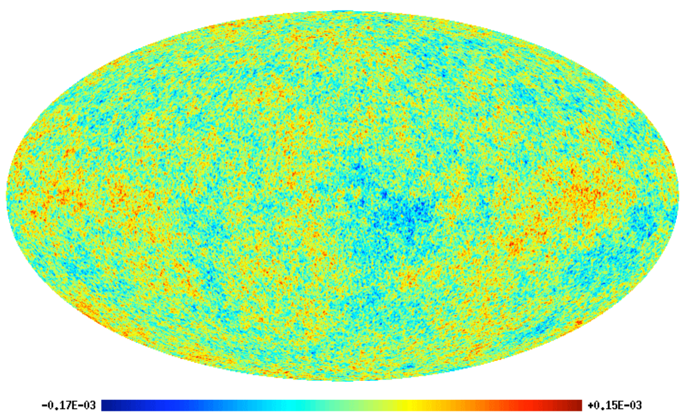
GTC: Performance

P	Part/ Cell	Power3 Seaborg		Itanium2 Thunder		Opteron Jacquard		X1 Phoenix		X1E Phoenix		SX6 ES		SX8 HLRS	
		GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk
128	200	0.14	9%	0.39	7%	0.59	13%	1.2	9%	1.7	10%	1.9	23%	2.3	14%
256	400	0.14	9%	0.39	7%	0.57	13%	1.2	9%	1.7	10%	1.8	22%	2.3	15%
512	800	0.14	9%	0.38	7%	0.51	12%			1.7	9%	1.8	22%		
1024	1600	0.14	9%	0.37	7%							1.8	22%		

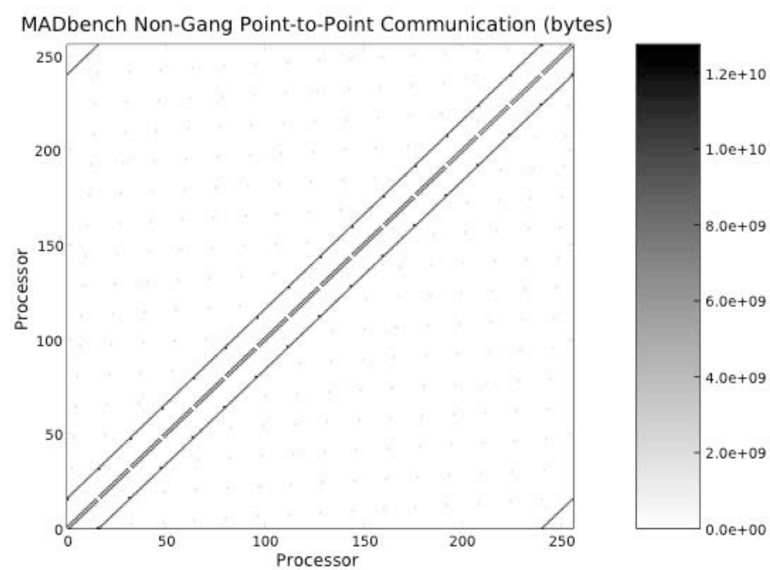
- New particle decomposition method to efficiently utilize large numbers of processors (as opposed to 64 on ES)
- Breakthrough of Tflop barrier on ES for important SciDAC code
  - 3.7 Tflop/s on 2048 processors
  - SX8 highest raw performance (ever) but lower efficiency than ES
- Opens possibility of new set of high-phase space-resolution simulations, that have not been possible to date
- X1 suffers from overhead of scalar code portions
- Scalar architectures suffer from low computational intensity, irregular data access, and register spilling
- Opteron/IB is 50% faster than Itanium2/Quadrics and only 1/2 speed of X1
  - Opteron: on-chip memory controller and caching of FP L1 data
- Original (unmodified) X1 version performed 12% *slower* on X1E
  - Recent additional optimizations increased performance by 50%!
- Chosen as HPCS benchmark

Cosmology: MADCAP

- Anisotropy Dataset Computational Analysis Package
- Optimal general algorithm for extracting key cosmological data from Cosmic Microwave Background Radiation (CMB)
- Anisotropies in the CMB contains early history of the Universe
- Recasts problem in dense linear algebra: ScaLAPACK
- Out of core calculation: holds approx 3 of the 50 matrices in memory



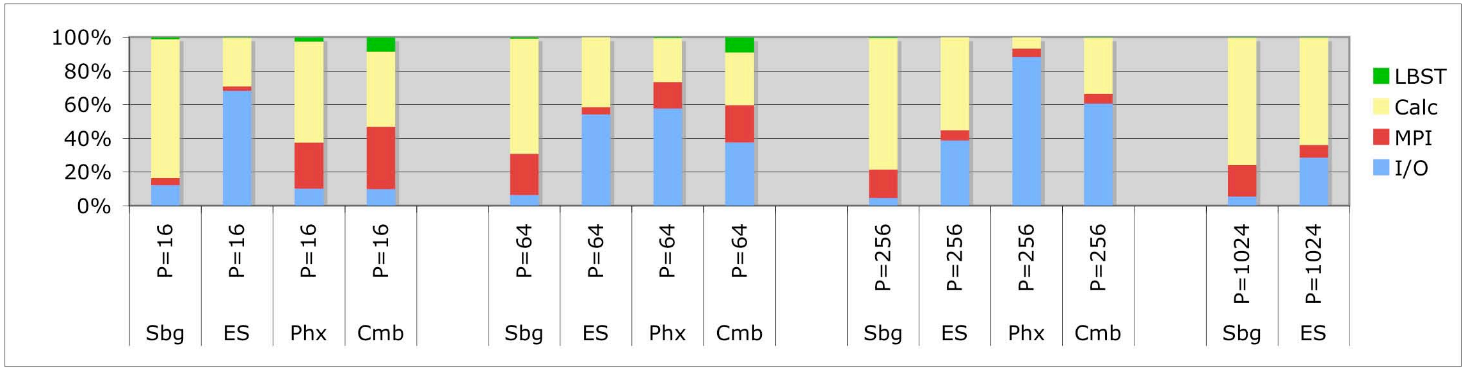
Temperature anisotropies in CMB (Boomerang)



Developed by Julian Borrill, LBNL

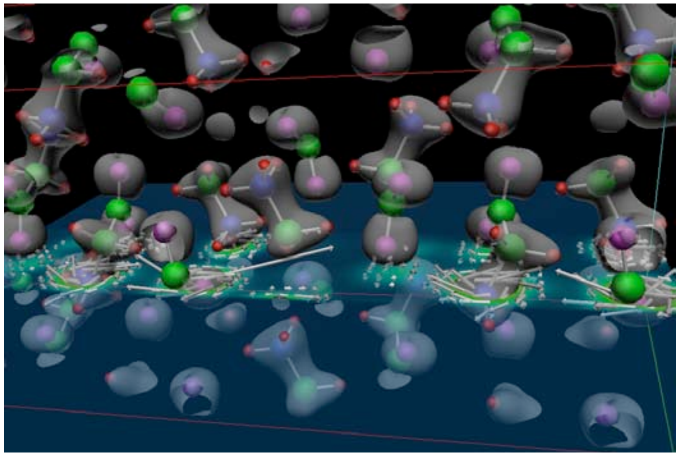
MADCAP: Performance

Number Pixels	P	NERSC (Power3)		Columbia (Itnm2)		Phoenix (X1)		ES (SX6*)	
		GFs/P	%pk	GFs/P	%pk	GFs/P	%pk	GFs/P	%pk
10K	64	0.73	49%	1.2	20%	2.2	17%	2.9	37%
20K	256	0.76	51%	1.1	19%	0.6	5%	4.0	50%
40K	1024	0.75	50%					4.6	58%



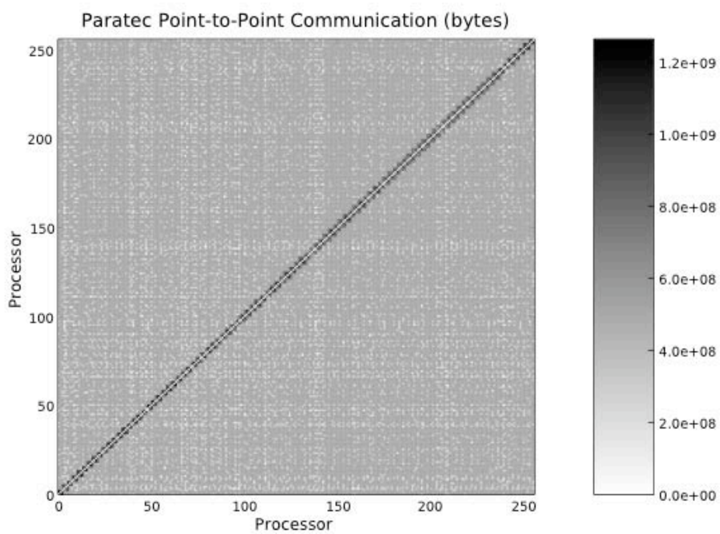
- Overall performance can be surprising low, for dense linear algebra code
- I/O takes a heavy toll on Phoenix and Columbia: I/O optimization in progress
- NERSC Power3 shows best system balance wrt to I/O
- ES lacks high-performance parallel I/O (code rewritten to use local disks)
- Developed MadBench benchmark with full complexity of application
- Starting collaboration with several groups including FastOS community

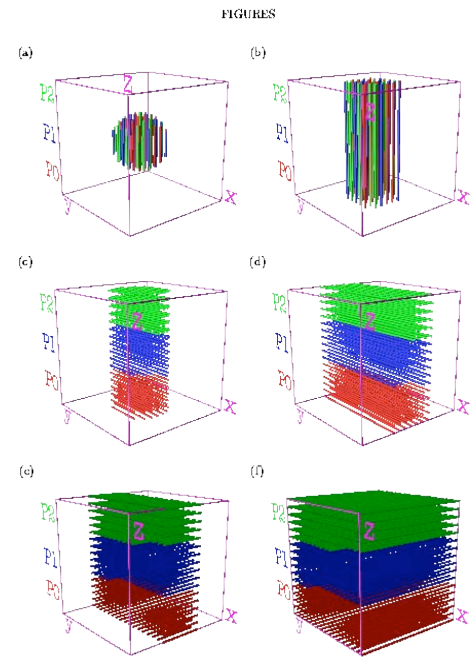
Material Science: PARATEC



Crystallized glycine induced current & charge

- PARATEC performs first-principles quantum mechanical total energy calculation using pseudopotentials & plane wave basis set
- Density Functional Theory to calc structure & electronic properties of new materials
- *DFT calc are one of the largest consumers of supercomputer cycles in the world*
- 33% 3D FFT, 33% BLAS3, 33% Hand coded F90
- Part of calculation in real space other in Fourier space
  - Uses specialized 3D FFT to transform wavefunction





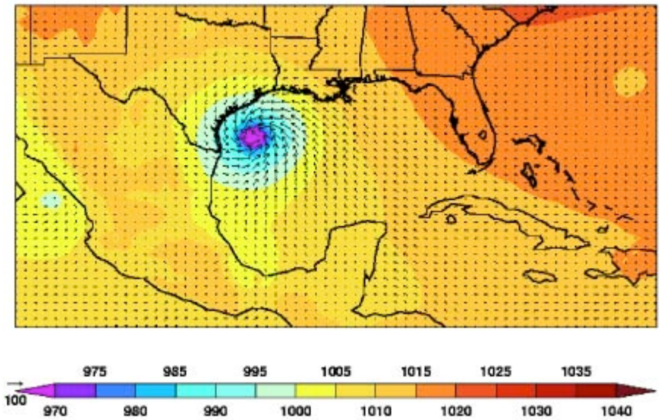
PARATEC: Performance

Problem	P	Power3 Seaborg GFs/P	Power3 Seaborg %pk	Itanium2 Thunder GFs/P	Itanium2 Thunder %pk	Opteron Jacquard GFs/P	Opteron Jacquard %pk	X1 Phoenix GFs/P	X1 Phoenix %pk	X1E Phoenix GF/s/P	X1E Phoenix %pk	SX6 ES GFs/P	SX6 ES %pk	SX8 HLRS GFs/P	SX8 HLRS %pk
488 Atom CdSe Quantum Dot	128	0.93	62%	2.8	51%			3.2	25%	3.8	21%	5.1	64%	7.5	64%
	256	0.85	67%	2.6	47%	2.0	45%	3.0	24%	3.3	18%	5.0	62%	6.8	62%
	512	0.73	49%	2.4	44%	1.0	22%			2.2	12%	4.4	55%		
	1024	0.60	40%	1.8	32%							3.6	46%		

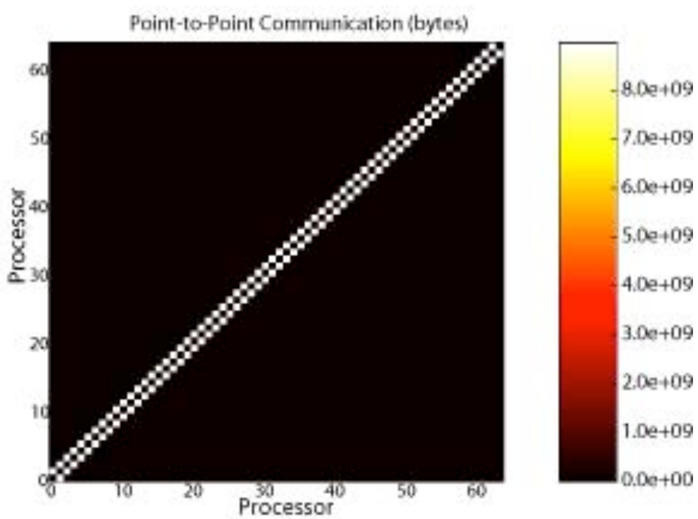
- All architectures generally achieve high performance due to computational intensity of code (BLAS3, FFT)
- ES achieves highest overall performance to date: 5.5Tflop/s on 2048 procs
  - Main ES advantage for this code is fast interconnect
  - Allows never before possible, high resolution simulations
  - Qdot: Largest cell-size atomistic experiment ever run using PARATEC
- SX8 achieves highest per-processor performance
- X1 shows lowest % of peak
  - Non-vectorizable code much more expensive on X1 (32:1)
  - Lower bisection bandwidth to computational ratio (2D Torus)
  - Performance is comparable to Itanium2

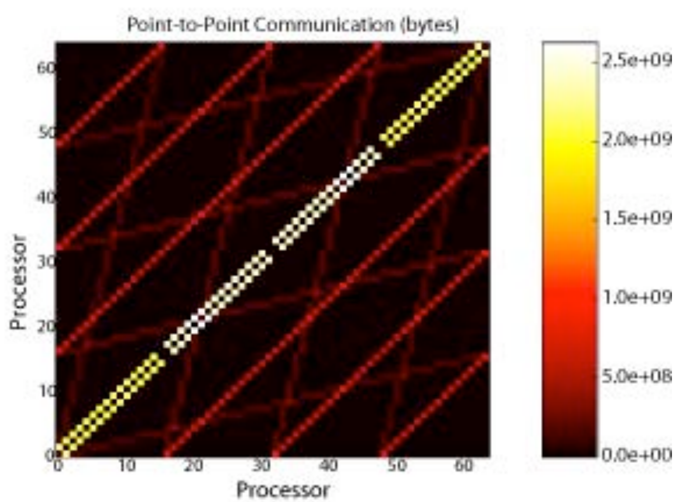
Developed by Andrew Canning with Louie and Cohen's groups (UCB, LBNL)

Climate: FVCAM



- Atmospheric component of CCSM
- AGCM: consists of physics (PS) and dynamical core (DC)
- DC approximates Navier-Stokes eqn's to describe dynamics of atmosphere
- PS: caculates source terms to equations of motion:
  - Turbulance, radiative transfer, clouds, etc
- Default approach uses spectral transform (1D decomp)
- Finite volume (FV) approach uses a 2D decomposition in latitude and level: allows higher concurrency
  - Requires remapping between Lagrangian surfaces and Eulerian reference frame

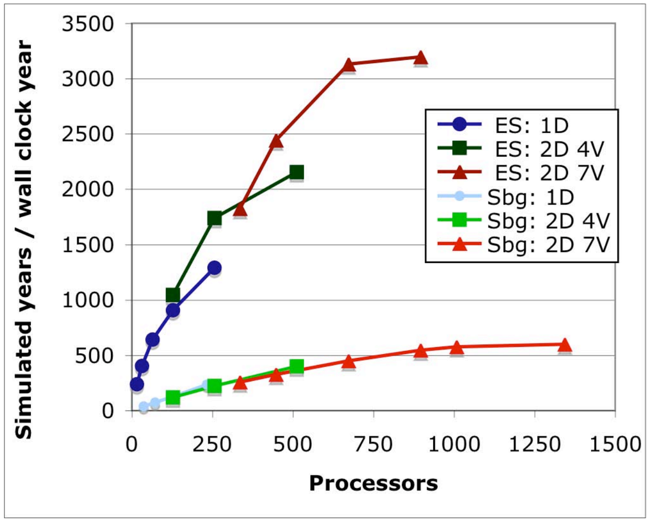




Experiments conducted by Michael Wehner, vectorized by Pat Worley, Art Mirin, Dave Parks

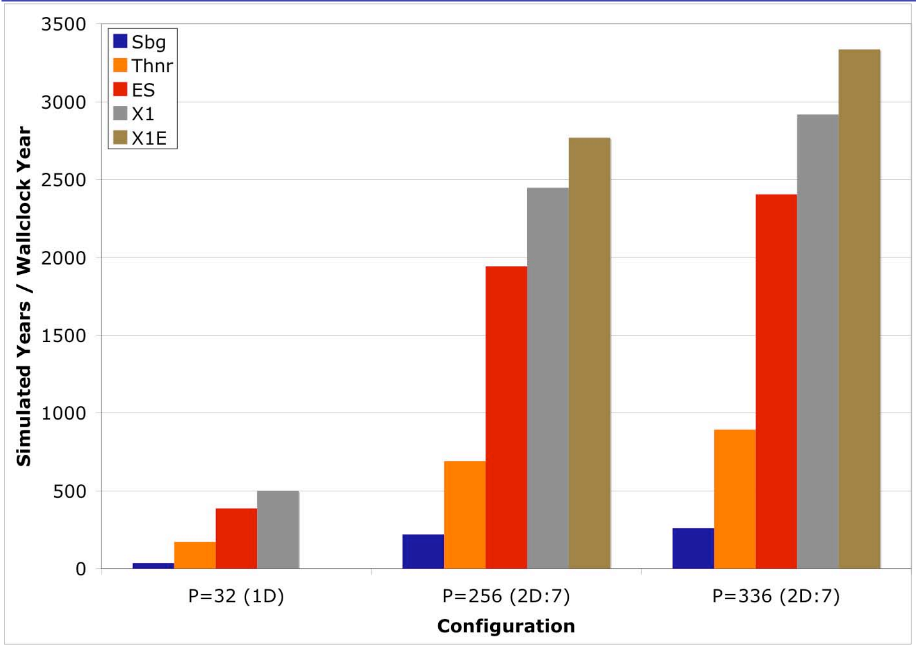
FVCAM3.0: Performance

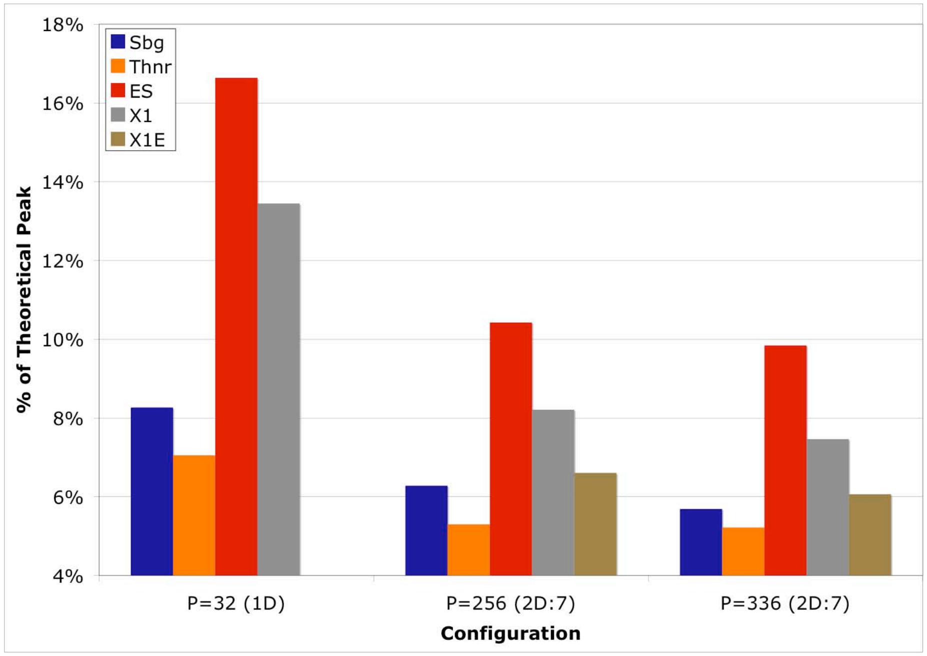
CAM3.0 results on ES and Power3, using *D* Mesh (0.5ºx0.625º)



- First published results showing high resolution vector performance
  - Requires multi-institution collaboration
- 2D approach allows both architectures to effectively use >2X as many procs
- At high concurrencies both platforms achieve low % peak (about 4%)
  - ES suffers from short vector lengths for fixed problem size, esp for FFTs
  - ES efficiency starts at 10% for small concurrency
- Increasing vertical discretizations (1,4,7) allows higher concurrencies
- ES can achieve more than 1000 simulation year / wall clock year (3200 on 896 processors), NERSC Power3 cannot exceed 600 regardless of concurrency
  - Speed up of 1000x or more is necessary for reasonable turnaround time
- Preliminary results: CAM3.1 experiments currently underway on ES, X1, Thunder, Power3

FVCAM3.1: Performance





- First comparison of X1E and ES
  - Results shown for latest version of FVCAM3.1
- Raw speed X1E: 1.14X X1, 1.4X ES, 3.7X Thunder, 13X Seaborg
- % of peak: ES 10%, X1 7.5%, X1E 6%, Seaborg 5.7%, Thunder 5.2%
- In-depth analysis and finer-grained resolution planned
- Collaborative effort for important SciDAC code: LBNL, LLNL, ORNL, ESC, NEC

Office of Science
U.S. DEPARTMENT OF ENERGY

Performance Overview

Code P=256	% Peak					Speedup ES vs			
	Pwr3	**Itnm2**	**X1**	**ES**	**SX8**	**Pwr3**	**Itan2**	**X1**	**SX8**
CACTUS	5%	10%	5%	34%	27%	38.6	4.9	4.0	0.6
LBMHD	9%	6%	41%	68%	60%	39.3	17.2	1.1	0.7
GTC	9%	7%	9%	20%	15%	11.4	4.1	1.3	0.7
MADCAP	51%	19%	5%	50%		5.3	3.6	6.7	
PARATEC	57%	47%	24%	62%	62%	5.9	1.9	1.7	0.7
FVCAM	6%	6%	8%	10%		8.9	2.5	0.7	
Average	*23%*	*14%*	*15%*	*41%*	*41%*	*18.2*	*5.7*	*2.6*	*0.7*

- Work fosters diverse collaborations and new optimization techniques
  - HPCS:Cache Oblivious, SciDAC: GTC particle decomp, FastOS I/O optimization
- Tremendous potential of vector architectures:
  - Vector systems allows resolution not possible with scalar platforms
  - Opportunity to perform scientific runs at unprecedented scale
- Evaluation codes contain sufficient regularity in computation for high vector performance
  - *Much more difficult to evaluate codes poorly suited for vectorization*
- Vectors potentially at odds w/ emerging techniques (irregular, multi-physics, multi-scale)
- Plan to expand scope of application domains/methods:
  - Build on existing code base and collaborative efforts
  - Sparse Methods, AMR, Life Sciences
- Next step latest HEC platforms with focus on ultra-parallel systems (BG/L)

BIPS

Future Plans

- Continue investigating vector performance but shift focus to ultra-scale architectures, network degree and level of integration
  - How efficient are ultra-scale low-power machines for DOE applications?
  - Under what circumstances can low-degree networks be used effectively?
  - Which codes benefit from tight network integration (low latency, SAS) ?
  - Given limitations of single processor scaling: what types of fine grained (on-chip) parallelism is most effective for scientific apps?
  - How do memory system designs (cache,cachless, cache incoherent) affect application performance?
  - What is value of shared memory hardware (e.g. CC-NUMA of Columbia)?
- Leverage existing application expertise and performance data
- Evaluate more complex irregular algorithms: AMR, sparse, particle
- Examine leading HPC platforms
  - BG/*, SX-8, X1E, X2, Columbia, Power5, Thunder, XT3, XD1
- Interested in exploring performance MPI alternatives (CAF, UPC)
- Perform in depth application characterizations
- Continue collaborations effort with HPCS, FastOS, PERC, SciDAC

Publications

- L. Oliker, J. Carter, M. Wehner, A. Canning, S. Ethier, B. Govindasamy, A. Mirin, D. Parks, P. Worley, "Performance of Ultra-Scale Applications on Leading Vector and Scalar HPC Platforms", **SC 2005**
- L. Oliker, A. Canning, J. Carter, J. Shalf, and S. Ethier. "*Scientific Computations on Modern Parallel Vector Systems*", **SC 2004** *****Nominated Best Paper award*
- L. Oliker, J. Carter, J. Shalf, D. Skinner, S. Ethier, R. Biswas, J. Djomehri, R. Van der Wijngaart. "*Evaluation of Cache-based Superscalar and Cacheless Vector Architectures for Scientific Computations*", **SC 2003**
- J. Borrill, J. Carter, D. Skinner, L. Oliker, R. Biswas ,"*Integrated Performance Monitoring of a Cosmology Application on Leading HEC Platforms*." ICPP2005 *Nominated for Best Paper award*
- J. Carter, J. Borrill, and L. Oliker. "*Performance Characteristics of a Cosmology Package on Leading HPC Architectures*", International Conference on Higher Performance Computing: HIPC 2004 *Nominated for Best Paper award*
- L. Oliker, R. Biswas, Rob Van der Wijngaart, David Bailey, Allan Snavely, "*Performance Evaluation and Modeling of Ultra-Scale Systems*", SIAM Publications Frontiers of Parallel Processing for Scientific Computing, to appear
- L. Oliker, A. Canning, J. Carter, J. Shalf, et al "*Ultra-scale Applications on Leading Vector and Scalar HPC Systems*", Journal of the Earth Simulator, 2005.
- L. Oliker, J. Carter, J. Shalf, D. Skinner, S. Ethier, R. Biswas, J. Djomehri, R. Van der Wijngaar "*Performance Evaluation of the SX-6 Vector Architecture for Scientific Computations*", Concurrency & Computation: Practice & Experience 2005
- Horst Simon, et al "*Science Driven System Architecture: A New Process for Leadership Class Computing*", Journal of the Earth Simulator, 2005.
- L. Oliker and R. Biswas, "*Performance Modeling and Evaluation of Ultra-Scale Systems*", Minisymposium organized a SIAM Conference on Parallel Processing for Scientific Computing: SIAMPP 2004.
- L. Oliker, J. Borrill, A. Canning, J. Carter, H. Shan, D. Skinner, R. Biswas, J. Djomheri, "*A Performance Evaluation of the Cray X1 for Scientific Applications*", VECPAR 2004.
- H. Shan, E. Strohmaier, L. Oliker, "*Optimizing Performance of Superscalar Codes For a Single Cray X1 MSP Processor*", 46th Cray User Group Conference, CUG 2004.
- G. Griem, L. Oliker, J. Shalf, K. Yelick, "*Identifying Performance Bottlenecks on Modern Microarchitectures using an Adaptable*
- *Probe*", Performance Modeling, Evaluation, Optimization of Parallel & Distributed Systems PMEO 2004

Collaborators

- Rupak Biswas, NASA Ames
- Andrew Canning LBNL
- Jonathan Carter, LBNL
- Stephane Ethier, PPPL
- Erich Strohmaier, LBNL
- Bala Govindasamy, LLNL
- Hongzhang Shan, LBNL
- Art Mirin, LLNL
- David Parks, NEC
- John Shalf, LBNL
- David Skinner, LBNL
- Yoshinori Tsuda, JAMSTEC
- Michael Welcome, LBNL
- Michael Wehner, LBNL
- Patrick Worley, ORNL